Contextual Recognition of Head Gestures

Louis-Philippe Morency
Computer Sciences and AI Lab., MIT
Cambridge, MA 02139, USA
lmorency@csail.mit.edu

Candace Sidner
Mitsubishi Electric Research Laboratories
Cambridge, MA 02139, USA
sidner@merl.com

Christopher Lee
Mitsubishi Electric Research Laboratories
Cambridge, MA 02139, USA
lee@merl.com

Trevor Darrell
Computer Sciences and AI Lab., MIT
Cambridge, MA 02139, USA
trevor@csail.mit.edu

ABSTRACT

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. We investigate how dialog context from an embodied conversational agent (ECA) can improve visual recognition of user gestures. We present a recognition framework which (1) extracts contextual features from an ECA's dialog manager, (2) computes a prediction of head nod and head shakes, and (3) integrates the contextual predictions with the visual observation of a vision-based head gesture recognizer. We found a subset of lexical, punctuation and timing features that are easily available in most ECA architectures and can be used to learn how to predict user feedback. Using a discriminative approach to contextual prediction and multi-modal integration, we were able to improve the performance of head gesture detection even when the topic of the test set was significantly different than the training set.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Motion*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Discourse*

General Terms

Algorithms, Languages

Keywords

Context-based recognition, Dialog context, Embodied conversational agent, Head gestures, Human-computer interaction

1. INTRODUCTION

During face-to-face conversation, people use visual feedback to communicate relevant information and to synchronize rhythm between participants. A good example of nonverbal feedback is head nodding and its use for visual grounding, turn-taking and answering yes/no questions. When recognizing visual feedback, people use more than their visual perception. Knowledge about the current topic and expectations from previous utterances help guide our visual perception in recognizing nonverbal cues. Our goal is to equip an embodied conversational agent (ECA) with the ability to use contextual information for performing visual feedback recognition much in the same way people do.

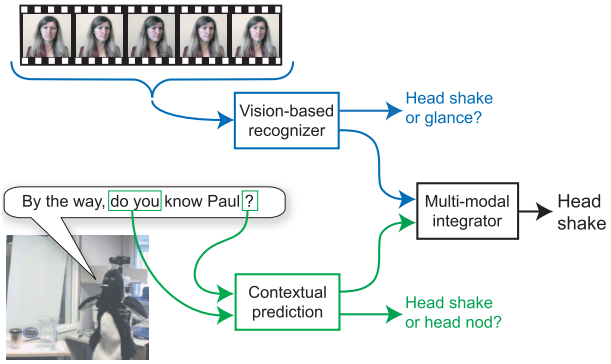

Figure 1: Contextual recognition of head gestures during face-to-face interaction with an embodied agent.

In the last decade, many ECAs have been developed for face-to-face interaction. A key component of these systems is the dialogue manager, usually consisting of a history of the past events, the current state, and an agenda of future actions. The dialogue manager uses contextual information to decide which verbal or nonverbal action the agent should perform next. This is called context-based synthesis.

Contextual information has proven useful for aiding speech recognition [9]. In [9], the grammar of the speech recognizer dynamically changes depending on the agent's previous action or utterance. In a similar fashion, we want to develop a context-based visual recognition module that builds upon the contextual information available in the dialogue manager to improve performance.

The use of dialogue context for visual gesture recognition has, to our knowledge, not been explored before for conversational interaction. In this paper we present a prediction framework for incorporating dialogue context with vision-based head gesture recognition.

ICMI'05, October 4–6, 2005, Trento, Italy.

The contextual features are derived from the utterances of the ECA, which is readily available from the dialogue manager. We highlight three types of contextual features: lexical, punctuation, and timing, and selected a subset for our experiment that were topic independent. We use a discriminative approach to predict head nods and head shakes from a small set of recorded interactions. We then combine the contextual predictions with a vision-based recognition algorithm based on the frequency pattern of the user's head motion. Our context-based recognition framework allows us to predict, for example, that in certain contexts a glance is not likely whereas a head shake or nod is (as in Figure 1), or that a head nod is not likely and a head nod misperceived by the vision system can be ignored.

The following section describes related work on gestures with ECAs. Section 3 describes the contextual information available in most embodied agent architectures. Section 4 shows how we automatically extract a subset of this context to compute lexical, punctuation, and timing features. Section 5 demonstrates how we use the contextual features to predict head nods and head shakes. Section 6 describes how we integrate the contextual prediction with the results from a vision-only head gesture recognizer. Finally, we describe our experiments, performed on 16 video recordings of human participants interacting with a robot.

2. RELATED WORK

There has been considerable work on gestures with ECAs. Bickmore and Cassell developed an ECA that exhibited many gestural capabilities to accompany its spoken conversation and could interpret spoken utterances from human users [1]. Sidner *et al.* have investigated how people interact with a humanoid robot [14]. They found that more than half their participants naturally nodded at the robot's conversational contributions even though the robot could not interpret head nods. Nakano *et al.* analyzed eye gaze and head nods in computer–human conversation and found that their subjects were aware of the lack of conversational feedback from the ECA [12]. They incorporated their results in an ECA that updated its dialogue state. Numerous other ECAs (e.g. [19, 3]) are exploring aspects of gestural behavior in human-ECA interactions. Physically embodied ECAs—for example, ARMAR II [5, 6] and Leo [2]–have also begun to incorporate the ability to perform articulated body tracking and recognize human gestures.

Head pose and gesture offer several key conversational grounding cues and are used extensively in face-to-face interaction among people. Stiefelhagen developed several successful systems for tracking face pose in meeting rooms and has shown that face pose is very useful for predicting turn-taking [16]. Takemae *et al.* also examined face pose in conversation and showed that if tracked accurately, face pose is useful in creating a video summary of a meeting [17]. Siracusa *et al.* developed a kiosk front end that uses head pose tracking to interpret who was talking to who in conversational setting [15]. The position and orientation of the head can be used to estimate head gaze which is a good estimate of a person's attention. When compared with eye gaze, head gaze can be more accurate when dealing with low resolution images and can be estimated over a larger range than eye gaze [11].

Kapoor and Picard presented a technique to recognize head nods and head shakes based on two Hidden Markov Models (HMMs) trained and tested using 2D coordinate results from an eye gaze tracker [8]. Fugie *et al.* also used HMMs to perform head nod recognition [7]. In their paper, they combined head gesture detection with prosodic recognition of Japanese spoken utterances to determine strongly positive, weak positive and negative responses to yes/no type utterances.

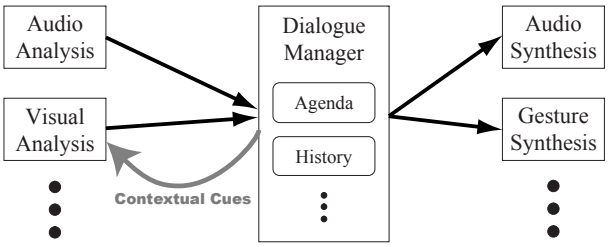

Figure 2: Simplified architecture for embodied conversational agent. Our method integrates contextual information from the dialogue manager inside the visual analysis module.

Context has been previously used in computer vision to disambiguate recognition of individual objects given the current overall scene category [18]. While some systems [12, 2] have incorporated tracking of fine motion actions or visual gesture, none have included top-down dialogue context as part of the visual recognition process.

3. DIALOG CONTEXT IN ECA ARCHITECTURE

During face-to-face interactions, people use knowledge about the current dialog to anticipate visual feedback from their interlocutor. As depicted in Figure 1, knowledge of the ECA's spoken utterance can help predict which visual feedback is most likely.

We use contextual information from the ECA's knowledge about the current dialog to improve recognition of visual feedback, more specifically, head gestures. The first step of this process is to determine which information already exists in most ECA architectures. Figure 2 shows our simplified architecture after analyzing several different systems [12, 13]. In this architecture, the dialogue manager contains two main sub-components, an agenda and a history[1]. The agenda keeps a list of all the possible actions the agent and the user (i.e. human participant) can do next. This list is updated by the dialogue manager based on its discourse model (prior knowledge) and on the history.

The simplified architecture depicted in Figure 2 highlights the fact that the dialog manager already processes contextual information in order to produce output for the speech and gesture synthesizer. The idea of this paper is to use this existing information to predict when visual feedback gestures from the user are likely. Since the dialog manager is already merging information from the input devices with the history and the discourse model, the output of the dialog manager will contain useful contextual information.

We highlight four types of contextual features easily available in the dialog manager:

LEXICAL FEATURES Lexical features are computed from the words said by the embodied agent. By analyzing the word content of the current or next utterance, one should be able to anticipate certain visual feedback. For example, if the current spoken utterance started with "Do you", the interlocutor will most likely answer using affirmation or negation. In this case, it is also likely to see visual feedback like a head nod or a head shake. On the other hand, if the current spoken utterance started with "What", then it's unlikely

[1]In our work we use the COLLAGEN conversation manager [13], but other dialogue managers provide these components as well.

to see the listener head shake or head nod–other visual feedback gestures (e.g., pointing) are more likely in this case.

PUNCTUATION FEATURES Punctuation features modify the way the text-to-speech engine will pronounce an utterance. Punctuation features can be seen as a substitute for more complex prosodic processing that are not yet available from most speech synthesizers. A comma in the middle of a sentence will produce a short pause, which will most likely trigger some feedback from the listener. A question mark at the end of the sentence represents a question that should be answered by the listener. When merged with lexical features, the punctuation features can help recognize situations (e.g., yes/no questions) where the listener will most likely use head gestures to answer.

TIMING Timing is an important part of spoken language and information about when a specific word is spoken or when a sentence ends is critical. This information can aid the ECA to anticipate visual grounding feedback. People naturally give visual feedback (e.g., head nods) during pauses of the speaker as well as just before the pause occurs. In natural language processing (NLP), lexical and syntactic features are predominant but for face-to-face interaction with an ECA, timing is also an important feature.

GESTURE DISPLAY Gesture synthesis is a key capability of ECAs and it can also be leveraged as a context cue for gesture interpretation. As described in [4], visual feedback synthesis can improve the engagement of the user with the ECA. The gestures expressed by the ECA influence the type of visual feedback from the human participant. For example, if the agent makes a deictic gesture, the user is more likely to look at the location that the ECA is pointing to.

The following section describes how we can automatically extract lexical, punctuation and timing features from the dialog system. As future work, we plan to experiment with a richer set of contextual cues including those based on gesture display.

4. CONTEXTUAL FEATURES

We want to automatically extract contextual information from the dialog manager rather than directly access the ECA internal state. Our proposed method extracts contextual features from the messages sent to the audio and gesture synthesizers. This strategy allows us to extract a summarized version of the dialog context while reducing the cost of extracting contextual cues. Since it does not presume any internal representation, our idea can be applied to most ECA architectures.

In our framework, the dialog manager sends a minimal set of information to the visual analysis module: the next spoken utterance, a time stamp and an approximated duration. The next spoken utterance contains the words, punctuation, and gesture information used to generate the ECA's actions. The utterance information is processed to extract the lexical, punctuation, timing, and gesture features described below. Approximate duration of utterances is generally computed by speech synthesizers and made available in the synthesizer API.

We extract bigrams and punctuation features from the spoken utterance. Bigrams (pairs of words that occur in close proximity to each other, and in particular order) are lexical features that can efficiently be computed given the transcript of the utterance. For example, given this utterance:

```
''Do you see the copper in the glass?''
```

the extracted bigrams would include: "do you", "you see", "see the", "the copper", "copper in", "in the", and "the glass". While a range of bigrams may be relevant to gesture context prediction, we currently focus on the single phrase "do you", as we observed it was an efficient predictor of a yes/no question in many of our training dialogs. Other bigram features will probably be useful as well, and could be learned using a feature selection algorithm from a set of candidate bigram features.

We extract bigrams from the utterance and set the following binary feature:

$$f_{\text{"do you"}} = \begin{cases} 1 & \text{if bigram "do you" is present} \\ 0 & \text{if bigram "do you" is not present} \end{cases}$$

The punctuation feature is coded similarly:

$$f_{?} = \begin{cases} 1 & \text{if the sentence ends with "?"} \\ 0 & \text{otherwise} \end{cases}$$

The timing contextual feature f_t represents proximity to the end of the utterance. The intuition is that verbal and non-verbal feedback are most likely at pauses and also just before the pause occurs. This feature can easily be computed given only two values: t_0, the utterance start-time, and δ_t, the estimated duration of the utterance. Given these two values for the current utterance, we can estimate f_t at time t using:

$$f_t(t) = \begin{cases} 1 - \left| \frac{t - t_0}{\delta_t} \right| & \text{if } t \leq t_0 + \delta_t \\ 0 & \text{if } t > t_0 + \delta_t \end{cases}$$

The contextual features are evaluated for every frame acquired by the visual analysis module (about 18Hz). The lexical and punctuation features are evaluated based on the current spoken utterance. The effect of an utterance starts when it starts to be spoken and ends after the pause following the utterance. The top three graphs of Figure 3 show how two sample utterances will be coded for the bigram "do you", the question mark and the timing feature.

We selected our features so that they are topic independent. This means that we should be able to learn how to predict head gesture from a small set of interactions and then use this knowledge on a new set of interactions with a a different topic discussed by the human participant and the robot. However, different classes of dialogs might have different key features, and ultimately these should be learned using a feature selection algorithm (this is a topic of future work).

5. CONTEXTUAL PREDICTION

In this section, we first describe our discriminative approach to learning the influence of contextual features on visual feedback. We learn automatically a likelihood measure of certain visual gestures given a subset of contextual features. Then, we present an experiment where we predict head nods and head shakes just from linguistic data.

Our prediction algorithm takes as input the contextual features and outputs a margin for each visual gesture. The margin is a scalar value representing how likely it is that a specific gesture happens. In our experiments, we focus on two head gestures: head nods and head shakes.

We are using a multi-class Support Vector Machine (SVM) to estimate the prediction of each visual gesture. The margin $m(x)$ of the feature vector x, created from the concatenation of the contextual features, can easily be computed given the learned set of support vectors x_i, the associated set of labels y_i and weights w_i, and the bias b:

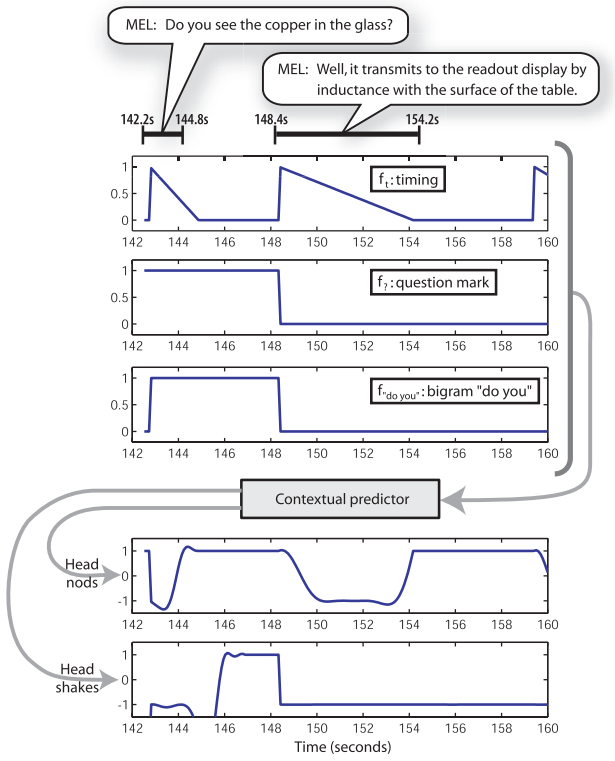

Figure 3: Prediction of head nods and head shakes based on contextual features: (1) distance to end-of-utterance when ECA is speaking, (2) type of utterance and (3) lexical bigram feature. We can see that the contextual predictor learned that head nods should happen near or at the end of an utterance or during a pause while head shakes are most likely at the end of a question.

$$m(x) = \sum_{i=1}^{l} y_i w_i K(x_i, x) + b \quad (1)$$

where l is the number of support vectors and $K(x_i, x)$ is the kernel function. In our experiments, we used a radial basis function (RBF) kernel:

$$K(x_i, x) = e^{-\gamma \|x_i - x\|^2} \quad (2)$$

where γ is the kernel smoothing parameter learned automatically using cross-validation on our training set. After training the multi-class SVM, we can easily compute a margin for each class and use this scalar value as a prediction for each visual gesture.

We trained the contextual predictor using a data set of seven video sequences where human participants conversed with a humanoid robot. The robot's spoken utterances were automatically processed, as described in Section 4, to compute the contextual features. A total of 236 utterances were used to train the multi-class SVM of our contextual predictor. Positive and negative samples were selected from the same data set based on manual transcription of head nods and head shakes.

Figure 3 displays the output of each class of our contextual predictor for a sample dialogue segment between the robot and a human participant held out from the training data. Positive margins represent a high likelihood for the gesture. It is interesting to observe that the contextual predictor automatically learned that head nods are more likely to occur around the end of an utterance or during a pause, while head shakes are most likely to occur after the completion of an utterance. More interestingly, it also learned that head shakes are directly correlated with the type of utterance (a head shake will most likely follow a question), and that head nods can happen at the end of a question to represent an affirmative answer and can also happen at the end of a normal statement to ground the spoken utterance.

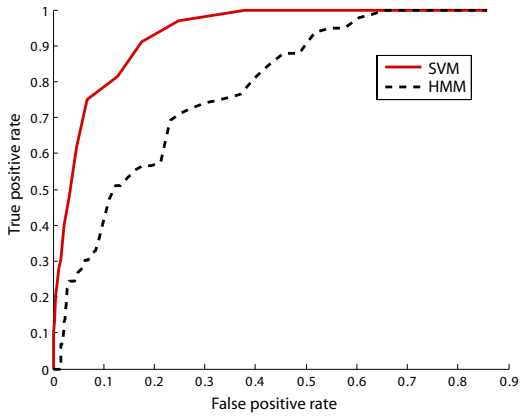

Figure 4: ROC curves for head nod recognition. Comparison of our SVM technique with a previously published HMM approach.

6. MULTI-MODAL INTEGRATION AND RECOGNITION

Having described how we anticipate a listener's visual feedback based on contextual information from an embodied conversational agent, we now integrate these predictions with observations from a vision-based head gesture recognizer. We will first describe the visual recognizer used during our experiments and then describe integration of contextual predictions.

6.1 Vision-based Head Gesture Recognition

We use a two-step process to recognize head gestures: we first track head position and rotation, and then use a computed head velocity feature vector to recognize head gestures. We use a head tracking framework that merges differential tracking with view-based tracking based on the system described by [10]. We found this tracker was able to track subtle movements of the head for a long period of time. While the tracker recovers the full 3-D position and velocity of the head, we found features based on angular velocities were sufficient for gesture recognition.

For vision-based gesture recognition (without dialog context), we trained a multi-class SVM with two different classes: head nods and head shakes. The head pose tracker outputs a head rotation velocity vector at each time step (sampled at approximately 18Hz). We transform the velocity signal into a frequency-based feature by applying a windowed FFT to each dimension of the velocity independently. We resample the velocity vector to have 32 samples per second. This transforms the time-based signal into an instantaneous frequency feature vector more appropriate for discriminative training. The multi-class SVM was trained using the RBF kernel described in Equation 2.

To evaluate our context-independent gesture recognition module, we compared it to previously published techniques using Hid-

den Markov Models (HMMs) [8, 7]. For the HMM technique, we used a variant of [8] based on head pose rather than eye gaze. We trained both classifiers (SVM and HMM) using a sampling of natural gestures and command-style gestures. Ten sequences were natural head gestures taken from interactions with an embodied agent, and 11 sequences were on-demand head gestures. The rotational velocity estimated by the head tracker was segmented manually to create two training data sets: head nods and head shakes. A third data-set (extra negative examples) was created from three minutes of video, where the subject is asked to move his/her head without producing any head nod or head shake gestures.

We tested our prototype on 30 video recordings of human participants interacting with an interactive robot (see Figure 5) that were not used in training. During these interactions, the robot spoke 935 utterances. Figure 4 shows the ROC (Receiver Operator Characteristic) curves of head nod detection for each technique. We can see that the SVM approach outperformed the HMM technique.

Based on these results, we decided to adopt the SVM approach for visual head gesture recognition. However, other classification schemes could also fit into our context-based recognition framework; all that we require for the multi-modal context fusion described below is that the vision-based head gesture recognizer return a single detection per head gesture. These detections are margins computed directly from the output of the multi-class SVM using Equation 1.

6.2 Integrated Recognition Framework

To recognize visual gestures in the context of the current dialog state, we fuse the output of the context predictor with the output of visual head gesture recognizer.

We considered two possible fusion schemes: (1) late fusion, where all context predictions are made independently of the visual observation and then merged together in a second step, and (2) early fusion, where predictions are made based on both the contextual features and the output from the vision-based recognizer. Limited initial experiments with both approaches suggested equivalent performance, so we selected late fusion because data acquisition for the contextual predictor is greatly simplified with this approach. Most recorded interactions between human participants and conversational robots do not include estimated head position. Since most natural language processing (NLP) learning methods work better with a large data set, a late fusion framework gives us the opportunity to train the contextual predictor on a larger data set of linguistic features.

Our integration component takes as input the margins from the contextual predictor (see Section 5) and the visual observations from the vision-based head gesture recognizer (Section 6.1), and recognizes if a head gesture has been expressed by the human participant. The output from the integrator is further sent to the dialog manager so it can be used to decide the next action of the ECA.

We use a multi-class SVM for the integrator since experimentally it gave us better performance than a linear classifier or simple thresholding. As mentioned earlier, the integrator could be trained on a smaller data set than the contextual predictor. However in our experiments, we trained the integrator on the same data set as the contextual predictor since our training data set included results from the head pose tracker. (Test data was withheld from both during evaluation.)

7. EXPERIMENTAL SETUP

The following experiment demonstrates how contextual features inferred from an agent's spoken dialogue can improve head nod and head shake recognition. The experiment compares the performance of the vision-only recognizer with the context-only prediction and with multi-modal integration.

Figure 5: Mel, the interactive robot, can present the iGlassware demo (table and copper cup on its right) or talk about its own dialog and sensorimotor abilities.

For this experiment, a first data set was used to train the contextual predictor and the multi-modal integrator (the same data set as described in Section 5), while a second data set with a different topic was used to evaluate the head gesture recognition performance. In the training data set, the robot interacted with the participant by demonstrating its own abilities and characteristics. This data set, called Self, contains 7 interactions. The test data set, called iGlass, consists of nine interactions of the robot describing the iGlassware invention ($\sim$340 utterances).

For both data sets, human participants were video recorded while interacting with the robot (see Figure 5). The vision-based head tracking and head gesture recognition was run online ($\sim$18Hz). The robot's conversational model, based on COLLAGEN [13], determines the next activity on the agenda using a predefined set of engagement rules, originally based on human–human interaction [14]. Each interaction lasted between 2 and 5 minutes.

During each interaction, we also recorded the results of the vision-based head gesture recognizer (described in Section 6.1) as well as the contextual cues (spoken utterances with start time and duration) from the dialog manager. These contextual cues were later automatically processed to create the contextual features (see Section 4) necessary for the contextual predictor (see Section 5).

For ground truth, we hand labeled each video sequence to determine exactly when the participant nodded or shook his/her head. A total of 274 head nods and 14 head shakes were naturally performed by the participants while interacting with the robot.

8. RESULTS

Our hypothesis was that the inclusion of contextual information within the head gesture recognizer would increase the number of recognized head nods while reducing the number of false detections. We tested three different configurations: (1) using the vision-only approach, (2) using only the contextual information as input (contextual predictor), and (3) combining the contextual information with the results of the visual approach (multi-modal integration).

Figure 6 shows head nod detection results for all 9 subjects used during testing. The ROC curves present the detection performance

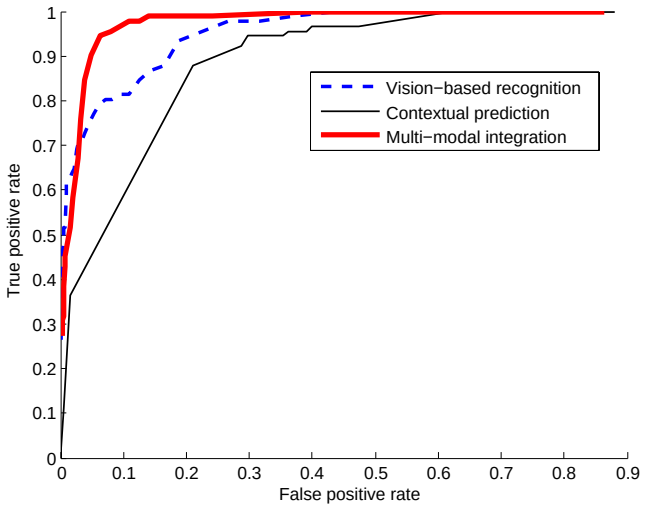

Figure 6: Head nod recognition curves when varying the detection threshold.

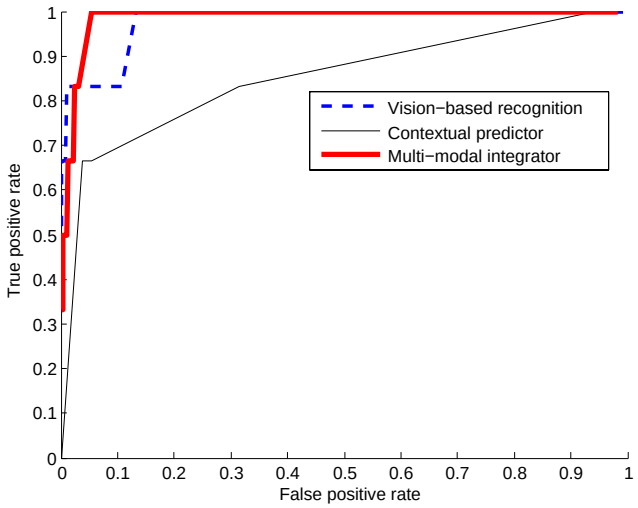

Figure 7: Head shake recognition curves when varying the detection threshold.

each recognition algorithm when varying the detection threshold. The areas under the curve for each techniques are 0.9543 for the vision only, 0.7923 for the predictor and 0.9722 for the integrator.

Figure 7 shows head shake detection results for each recognition algorithm when varying the detection threshold. The areas under the curve for each techniques are 0.9782 for the vision only, 0.8521 for the predictor and 0.9684 for the integrator.

Table 1 summarizes the results from Figures 6 and 7 by computing the true positive rates for the fixed negative rate of 0.05. Using a standard analysis of variance (ANOVA) on all the subjects, results on the head nod detection task showed a significant difference among the means of the 3 methods of detection: $F(2, 8) = 20.22$, $p = 0.002$, $d = 0.97$. Pairwise comparisons show a significant difference between all pairs, with $p = 0.006$, $p = 0.039$, and $p < 0.001$ for vision-predictor, vision-integrator, and predictor-integrator respectively. A larger number of samples would be necessary to see the same significance in head shakes.

We computed the true positive rate using the following ratio:

$$\text{True positive rate} = \frac{\text{Number of detected gestures}}{\text{Total number of ground truth gestures}}$$

A head gesture is tagged as detected if the detector triggered at least once during a time window around the gesture. The time window starts when the gesture starts and ends k seconds after the gesture. The parameter k was empirically set to the maximum delay of the vision-based head gesture recognizer (1.0 second). For the iGlass dataset, the total numbers of ground truth gestures were 91 head nods and 6 head shakes.

The false positive rate is computed at a frame level:

$$\text{False positive rate} = \frac{\text{Number of falsely detected frames}}{\text{Total number of non-gesture frames}}$$

A frame is tagged as falsely detected if the head gesture recognizer triggers and if this frame is outside any time window of a ground truth head gesture. The denominator is the total of frames outside any time window. For the iGlass dataset, the total number of non-gestures frames was 18246 frames and the total number of frames for all 9 interactions was 20672 frames.

	Vision	Predictor	Integrator
Head nods	75%	42%	**90%**
Head shakes	84%	67%	**100%**

Table 1: True detection rates for a fix false positive rate of 0.05.

Figure 8 shows the head nod recognition results for a sample dialogue. When only vision is used for recognition, the algorithm makes a mistake at around 101 seconds by detecting a false head nod. Visual grounding is less likely during the middle of an utterance. By incorporating the contextual information, our context-based gesture recognition algorithm is able to reduce the number of false positives. In Figure 8 the likelihood of a false head nod happening is reduced.

9. CONCLUSION AND FUTURE WORK

Our results show that contextual information can improve user gesture recognition for interactions with embodied conversational agents. We presented a prediction framework that extracts knowledge from the spoken dialogue of an embodied agent to predict which head gesture is most likely. By using simple lexical, punctuation, and timing context features, we were able to improve the recognition rate of the vision-only head gesture recognizer from 75% to 90% for head nods and from 84% to 100% for head shakes. As future work, we plan to experiment with a richer set of contextual cues including those based on gesture display, and to incorporate general feature selection to our prediction framework so that a wide range of potential context features can be considered and the optimal set determined from a training corpus.

10. REFERENCES

[1] Tim Bickmore and Justine Cassell. *J. van Kuppevelt, L. Dybkjaer, and N. Bernsen (eds.), Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.

[2] Breazeal, Hoffman, and A. Lockerd. Teaching and working with robots as a collaboration. In *The Third International*

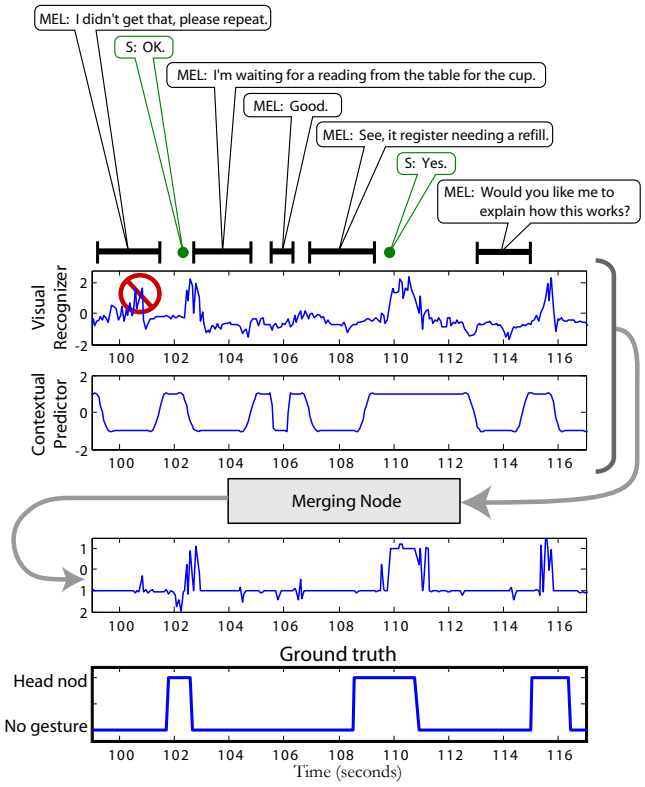

Figure 8: Head nod recognition results for a sample dialogue. The last graph displays the ground truth. We can observe at around 101 seconds (circled and crossed in the top graph) that the contextual information attenuates the effect of the false positive detection from the visual recognizer.

Conference on Autonomous Agents and Multi-Agent Systems AAMAS 2004, pages 1028–1035. ACM Press, July 2004.

[3] De Carolis, Pelachaud, Poggi, and F. de Rosis. Behavior planning for a reflexive agent. In *Proceedings of IJCAI*, Seattle, September 2001.

[4] Justine Cassell and Kristinn R. Thorisson. The poser of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 1999.

[5] Dillman, Becher, and P. Steinhaus. ARMAR II – a learning and cooperative multimodal humanoid robot system. *International Journal of Humanoid Robotics*, 1(1):143–155, 2004.

[6] Dillman, Ehrenmann, Steinhaus, Rogalla, and R. Zoellner. Human friendly programming of humanoid robots–the German Collaborative Research Center. In *The Third IARP Intenational Workshop on Humanoid and Human-Friendly Robotics*, Tsukuba Research Centre, Japan, December 2002.

[7] Shinya Fujie, Yasuhi Ejiri, Kei Nakajima, Yosuke Matsusaka, and Tetsunori Kobayashi. A conversation robot using head gesture recognition as para-linguistic information. In *Proceedings of 13th IEEE International Workshop on Robot and Human Communication, RO-MAN 2004*, pages 159–164, September 2004.

[8] A. Kapoor and R. Picard. A real-time head nod and shake detector. In *Proceedings from the Workshop on Perspective User Interfaces*, November 2001.

[9] Lemon, Gruenstein, and Stanley Peters. Collaborative activities and multi-tasking in dialogue systems. *Traitement Automatique des Langues (TAL), special issue on dialogue*, 43(2):131–154, 2002.

[10] Ali Rahimi Louis-Philippe Morency and Trevor Darrell. Adaptive view-based appearance model. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 803–810, 2003.

[11] L.-P. Morency, A. Rahimi, N. Checka, and T. Darrell. Fast stereo-based head tracking for interactive environment. In *Proceedings of the Int. Conference on Automatic Face and Gesture Recognition*, pages 375–380, 2002.

[12] Nakano, Reinstein, Stocky, and Justine Cassell. Towards a model of face-to-face grounding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.

[13] Rich, Sidner, and Neal Lesh. Collagen: Applying collaborative discourse theory to human–computer interaction. *AI Magazine, Special Issue on Intelligent User Interfaces*, 22(4):15–25, 2001.

[14] C. Sidner, C. Lee, C.D.Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1–2):140–164, August 2005.

[15] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell. Haptics and biometrics: A multimodal approach for determining speaker location and focus. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, November 2003.

[16] R. Stiefelhagen. Tracking focus of attention in meetings. In *Proceedings of International Conference on Multimodal Interfaces*, 2002.

[17] Y. Takemae, K. Otsuka, and N. Mukaua. Impact of video editing based on participants' gaze in multiparty conversation. In *Extended Abstract of CHI'04*, April 2004.

[18] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *IEEE Intl. Conference on Computer Vision (ICCV)*, Nice, France, October 2003.

[19] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual world. In *Proceedings of the International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2002)*, pages 766–773, July 2002.